KU LEUVEN

MLC 2014:
Regression and classification using random forest-based techniques

David Robben, Daan Christiaens, Frederik Maes, Paul Suetens

Medical Imaging Research Center, KU Leuven, Belgium

Introduction

- Goal
  - Comparison of supervised learning techniques
    - Used as a black box
  - Emphasis on correct estimation of performance

- Two problems
  - Binary classification
  - Regression

Our submission

- Three algorithms, all of them:
  - Tree-based
  - Used on both problems

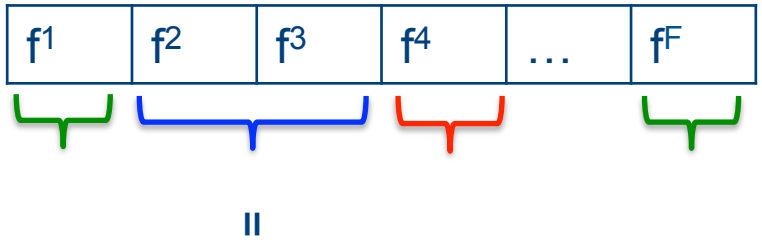
KU LEUVEN

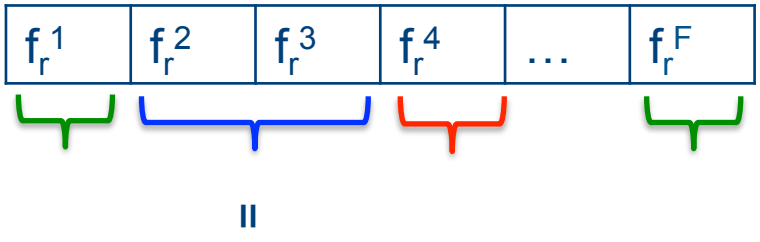
1. Random forests (RF)

- In my opinion, the default thing to try.
- Implementation from scikit-learn

2. Rotation forests

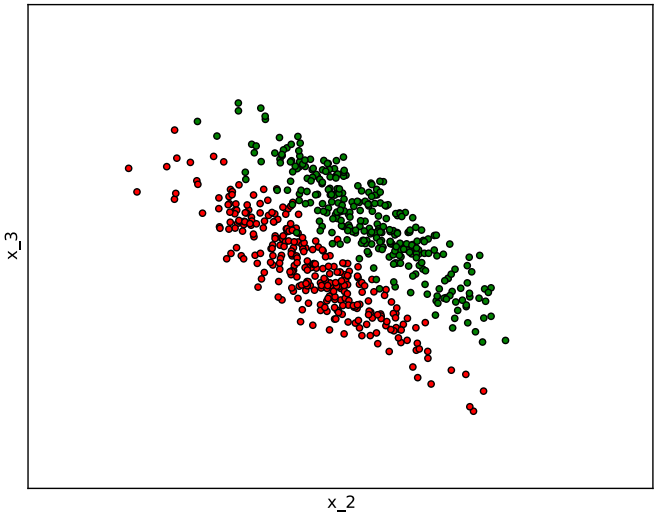
- Like random forests but
  - Each tree uses differently transformed features



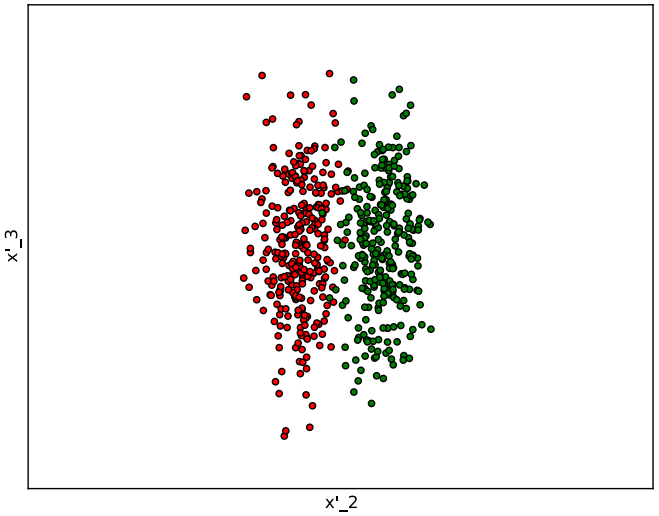


=

=

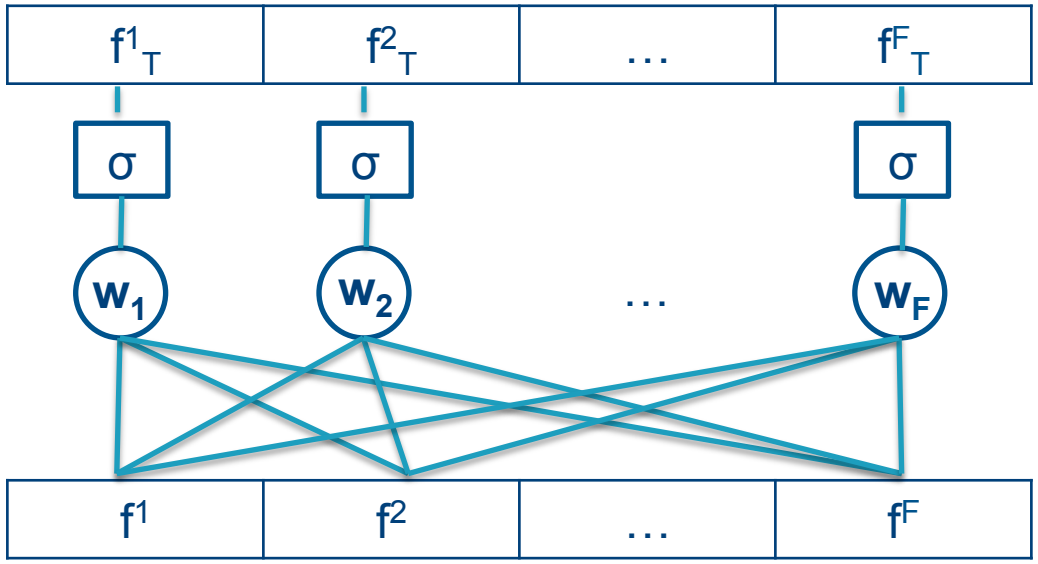


PCA



3. Denoising auto encoder followed by a RF

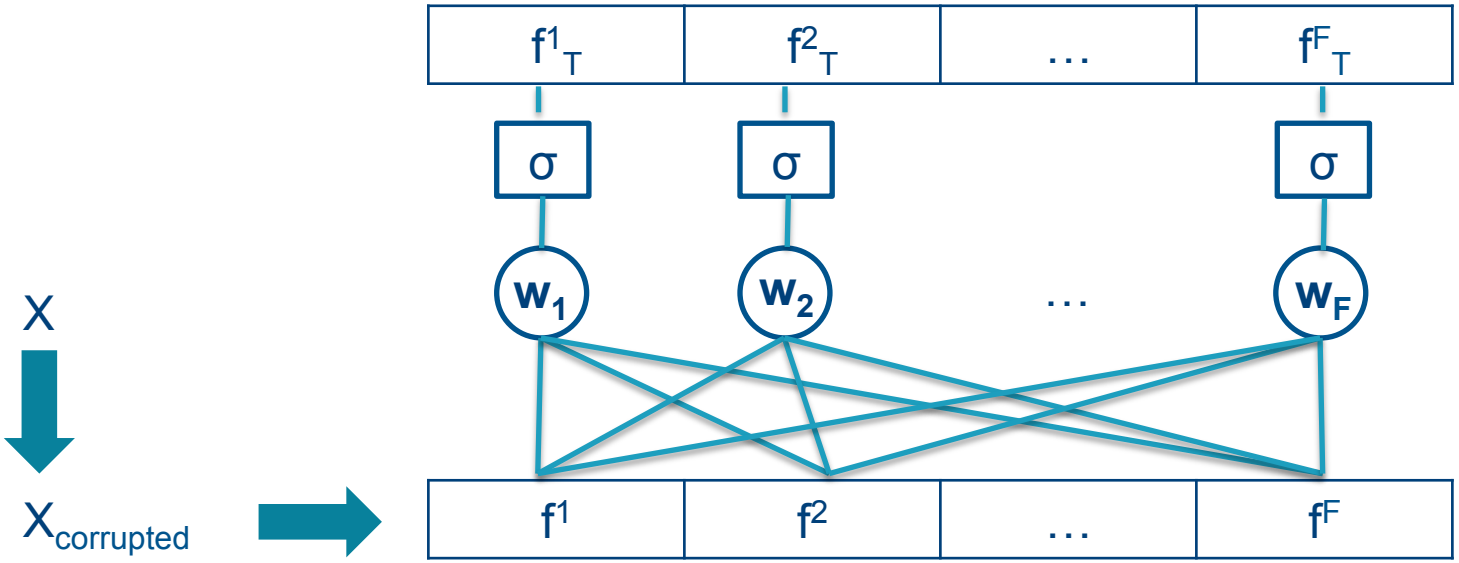
- Denoising auto encoder



KU LEUVEN

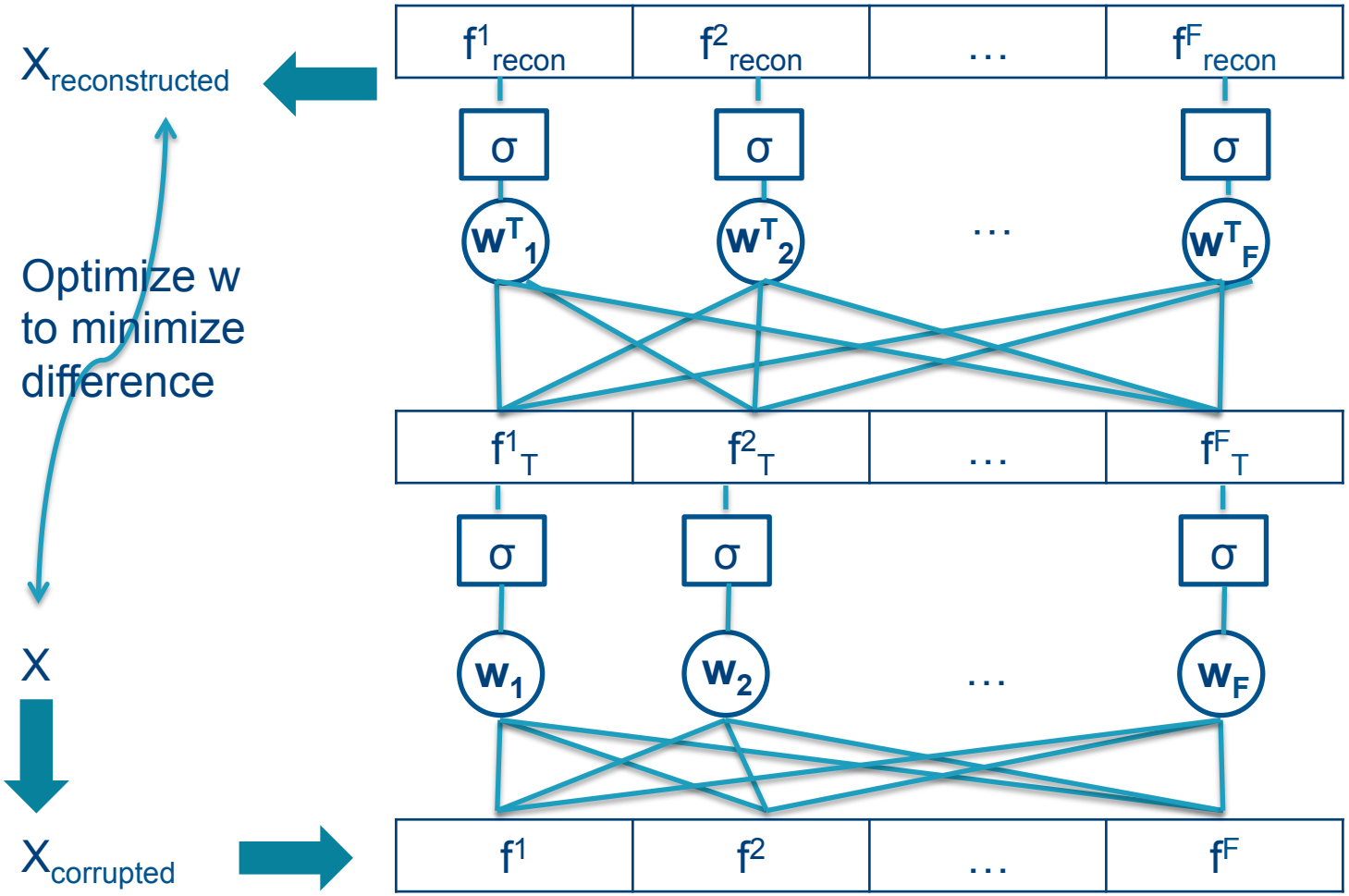
3. Denoising auto encoder followed by a RF

- Denoising auto encoder



3. Denoising auto encoder followed by a RF

- Denoising auto encoder



KU LEUVEN

3. Denoising auto encoder followed by a RF

- Denoising auto encoder
  - Unsupervised feature learning
  - ~ Robust non-linear PCA without dimensionality reduction
- Random forest on these features
- Implemented in Python, using scikit-learn and Theano, following deeplearning.net.

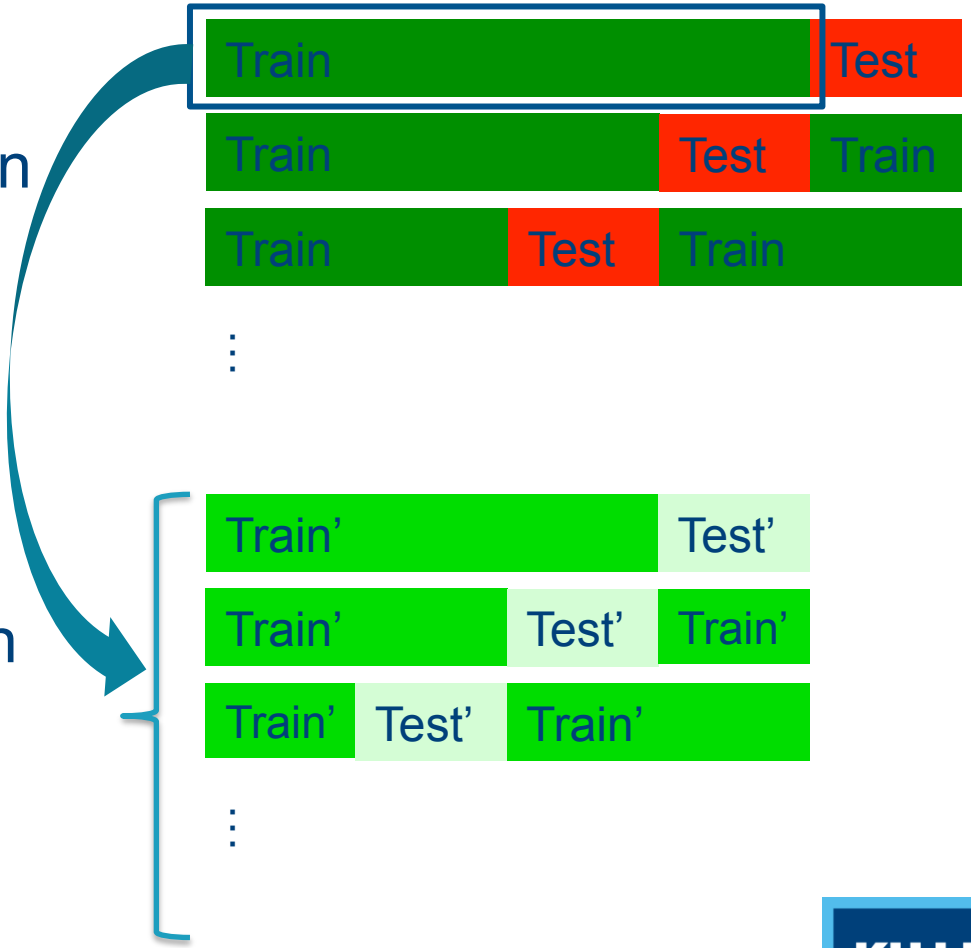
Evaluation using nested cross-validation

Outer cross-validation

Estimates performance of method

Inner cross-validation

Estimates performance for different hyper parameter settings



KU LEUVEN

Results

Classification accuracy	Random F	Rotation F	DAE + RF
Cross-validation	0.63	0.63	0.67
Test	0.62 (0.05)	0.60 (0.05)	0.50 (0.05)

#12

Classification AUC	Random F	Rotation F	DAE + RF
Cross-validation	0.68	0.70	0.72
Test	0.67 (0.05)	0.70 (0.05)	0.55 (0.06)

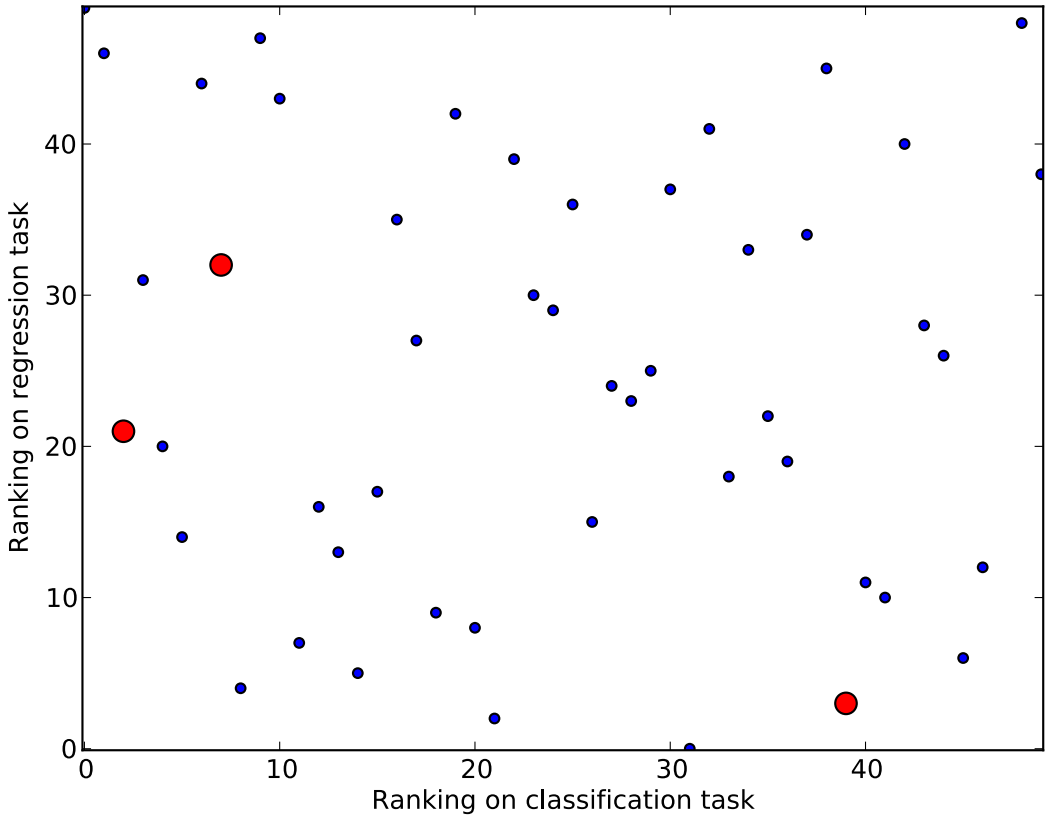
#3

Regression RMSE	Random F	Rotation F	DAE + RF
Cross-validation	0.88	0.93	0.87
Test	1.68 (0.10)	1.62 (0.09)	1.12 (0.07)

#5

KU LEUVEN

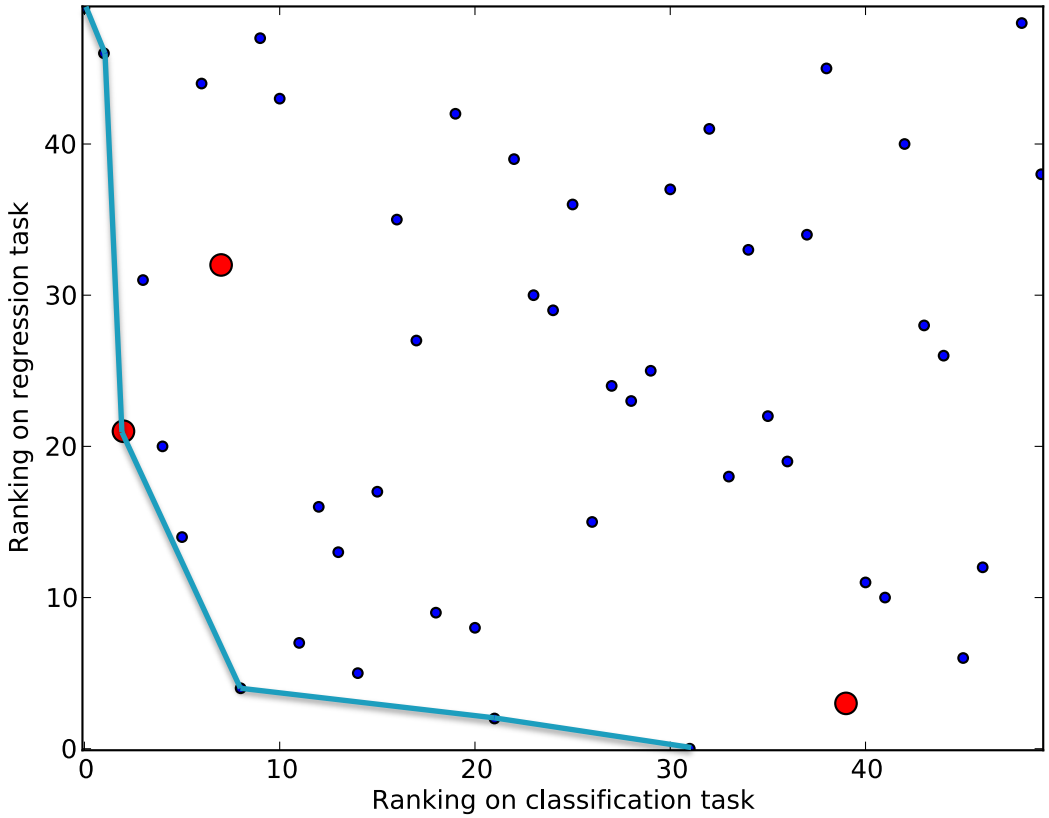
Discussion: no free lunch



Spearman rank correlation: -0.07

KU LEUVEN

Discussion: no free lunch



Pareto efficiency

KU LEUVEN

Conclusion

- Interesting challenge
- Difficult problem
- Estimation performance is hard
- No free lunch

KU LEUVEN

Questions?

KU LEUVEN

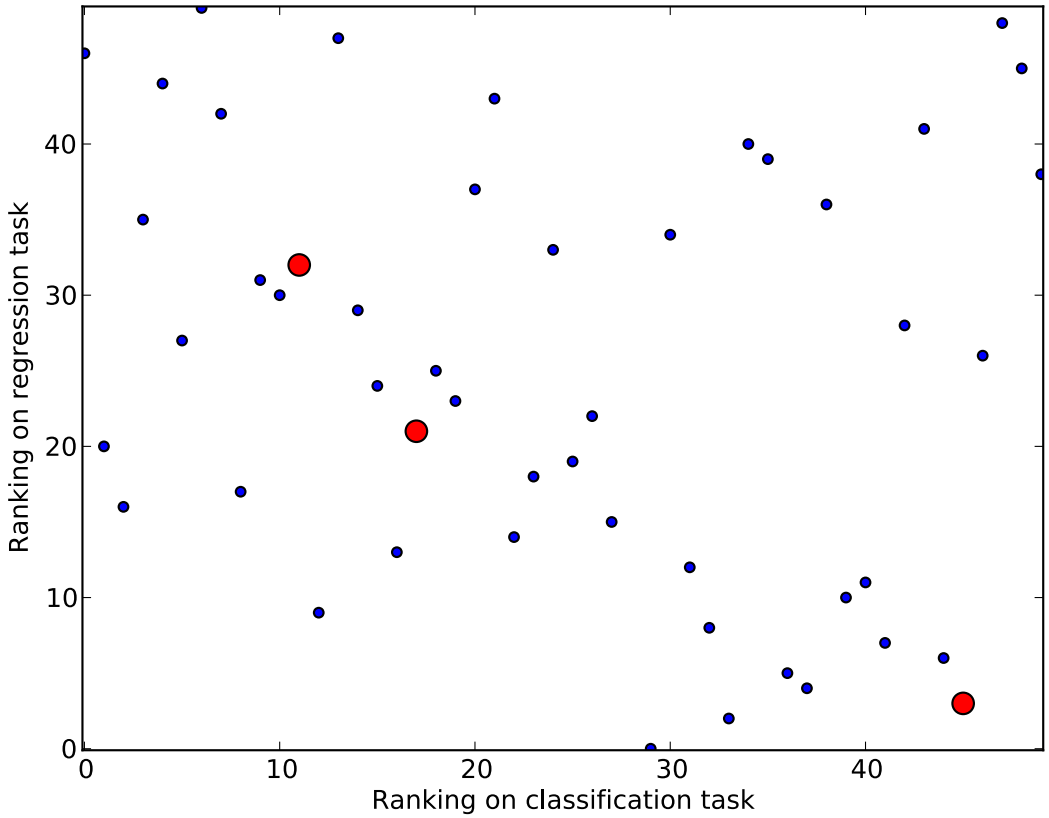
Hyper parameters and considered values

- For all three:
  - Minimum number of samples per leaf: 1,2,4,8
  - Number of trees: 500
  - Number of features considered when looking for best split: $\sqrt{}$(nb features)

- For the rotation forests:
  - PCA subset size: 4,8,16

- For the denoising auto encoder:
  - Learning rate: 0.1
  - Batch size: 20
  - Training epochs: 50
  - Corruption level: 0.1
  - Nb hidden nodes: 184 (=nb features)

Discussion: no free lunch



Spearman rank correlation: -0.21

KU LEUVEN

Results – classification

	Random F	Rotation F	DAE + RF
Accuracy – 5 fold	0.63	0.63	0.67
Accuracy – 50x 0.2 hold out	0.66	0.66	0.67
Accuracy – 100x 0.1 hold out	0.67	-	0.69
Accuracy – test	0.62 (0.05)	0.60 (0.05)	0.50 (0.05)

#12

	Random F	Rotation F	DAE + RF
AUC – 5 fold	0.68	0.70	0.72
AUC – 50x 0.2 hold out	0.75	0.73	0.67
AUC – 100x 0.1 hold out	0.71	-	0.69
AUC – test	0.67 (0.05)	0.70 (0.05)	0.55 (0.06)

#3

Results – regression

	Random F	Rotation F	DAE + RF
RMSE – 5 fold	0.88	0.93	0.87
RMSE – 50x 0.2 hold out	0.90	0.92	0.81
RMSE – 100x 0.1 hold out	0.84	-	0.82
RMSE - test	1.68 (0.10)	1.62 (0.09)	1.12 (0.07)

#5

Results – regression

	Random F	Rotation F	DAE + RF
RMSE – 5 fold	0.88	0.93	0.87
RMSE – 50x 0.2 hold out	0.90	0.92	0.81
RMSE – 100x 0.1 hold out	0.84	-	0.82
RMSE - test	1.68 (0.10)	1.62 (0.09)	1.12 (0.07)

#5

	Random F	Rotation F	DAE + RF
R– 5 fold	0.93	0.93	0.92
R– 50x 0.2 hold out	0.93	0.93	0.93
R– 100x 0.1 hold out	0.93	-	0.93
R- test	0.56 (0.07)	0.57 (0.07)	0.656 (0.07)